Métarecherche pour l'extraction lexicale bilingue à partir de corpus comparables

Amir Hazem, Emmanuel Morin, Sebastian Peña Saldarriaga

TALN 2011, Montpellier, 27 juin – 1er juillet 2011

Métarecherche pour l'extraction lexicale bilingue à partir de corpus comparables

Amir Hazem[1] Emmanuel Morin[1] Sebastián Peña Saldarriaga[2]
(1) Université de Nantes, LINA - UMR CNRS 6241
2 rue de la Houssinière, BP 92208, 44322 Nantes Cedex 03
(2) Synchromedia, École de technologie supérieure
1100 rue Notre-Dame Ouest, Montréal, Québec, Canada H3C 1K3
amir.hazem@univ-nantes.fr, emmanuel.morin@univ-nantes.fr, spena@synchromedia.ca

Résumé. Nous présentons dans cet article une nouvelle manière d'aborder le problème de l'acquisition automatique de paires de mots en relation de traduction à partir de corpus comparables. Nous décrivons tout d'abord les approches standard et par similarité interlangue traditionnellement dédiées à cette tâche. Nous réinterprétons ensuite la méthode par similarité interlangue et motivons un nouveau modèle pour reformuler cette approche inspirée par les métamoteurs de recherche d'information. Les résultats empiriques que nous obtenons montrent que les performances de notre modèle sont toujours supérieures à celles obtenues avec l'approche par similarité interlangue, mais aussi comme étant compétitives par rapport à l'approche standard.

Abstract. In this article we present a novel way of looking at the problem of automatic acquisition of pairs of translationally equivalent words from comparable corpora. We first describe the standard and extended approaches traditionally dedicated to this task. We then re-interpret the extended method, and motivate a novel model to reformulate this approach inspired by the metasearch engines in information retrieval. The empirical results show that performances of our model are always better than the baseline obtained with the extended approach and also competitive with the standard approach.

Mots-clés : Corpus comparables, lexiques bilingues, métarecherche.

Keywords: Comparable corpora, bilingual lexicon, metasearch.

1 Introduction

L'extraction de lexiques bilingues à partir de corpus comparables est un domaine de recherche en pleine effervescence qui vise notamment à offrir une alternative crédible à l'exploitation de corpus parallèles. En effet, les corpus parallèles sont par nature des ressources rares notamment pour les domaines spécialisés et pour des couples de langues ne faisant pas intervenir l'anglais, là ou les corpus comparables sont par essence des ressources abondantes puisque composés de documents partageant différentes caractéristiques telles que le domaine, le genre, la période, etc. sans être en correspondance de traduction. Les lexiques bilingues extraits à partir de corpus comparables sont néanmoins d'une qualité bien inférieure à ce qui peut être obtenu à partir de corpus parallèles. Cette difficulté à extraire des lexiques bilingues peu bruités à partir de corpus comparables explique pourquoi ce champ de recherche n'a pas encore franchi le cap de l'industrialisation à la différence des corpus parallèles et reste encore majoritairement cantonné à une activité de recherche prometteuse. La principale difficulté des approches liées à l'exploitation de corpus comparables par rapport aux corpus parallèles pour l'extraction de lexiques bilingues, est l'absence d'éléments d'ancrage entre les documents des langues source et cible composant le corpus comparable. Face à cette difficulté les différentes approches liées à l'exploitation de corpus comparables reposent sur la simple observation qu'un mot et sa traduction ont tendance à apparaître dans les mêmes environnements lexicaux. La mise en œuvre de cette observation repose sur l'identification d'*affinités du premier ordre* (*i.e.* identifier les mots qui sont susceptibles d'être trouvés dans le voisinage immédiat d'un mot donné) ou d'*affinités du second ordre* (*i.e.* identifier les mots qui partagent les mêmes environnements lexicaux sans nécessairement apparaître ensemble) (Grefenstette, 1994a, p. 279). Les approches associées à l'identification de ces affinités sont, d'une

part, l'approche standard (Rapp, 1995; Fung & McKeown, 1997) qui est l'approche majoritairement exploitée, et d'autre part, l'approche par similarité interlangue (Déjean & Gaussier, 2002).

Dans cette article, nous reprenons à notre compte l'idée de (Fung, 1998) qui indique que l'extraction de lexiques bilingues à partir de corpus comparables peut être approchée comme un problème de recherche d'information. Dans cette représentation, la requête serait alors les mots à traduire et les documents retournés par le moteur de recherche les candidats à la traduction de ce mot. Et de la même manière que les documents retournés sont ordonnés suivant leur adéquation avec la requête, les traductions candidates sont classées en fonction de leur pertinence par rapport au mot à traduire. Nous souhaitons donc poursuivre plus en avant cette analogie et proposer une amélioration significative à l'approche par similarité interlangue en considérant l'extraction de lexiques bilingues comme un problème de fusion de résultats analogue à celui rencontré par les métamoteurs de recherche d'information. Nous faisons ainsi l'hypothèse que le fait de combiner différentes sources d'information permet de renforcer globalement la méthode par similarité interlangue.

Dans la suite de cet article, nous commençons par rappeler en section 2 les deux méthodes phares en extraction de lexiques bilingues à partir de corpus comparables, à savoir les méthodes dites standard et par similarité interlangue. La section 3 est quant à elle dédiée à la présentation de notre approche par métarecherche qui revisite l'approche par similarité interlangue. La section 4 se concentre sur l'évaluation des trois méthodes mises en œuvre et ouvre une discussion sur les limites de notre approche. Enfin la section 5 vient conclure ce travail.

2 Principales approches en extraction lexicale bilingue à partir de corpus comparables

Dans cette section, nous allons décrire les deux principales approches dédiées à l'extraction de lexiques bilingues à partir de corpus comparables, à savoir : l'*approche standard*, puis l'*approche par similarité interlangue*.

2.1 Approche standard

Les principaux travaux en extraction de lexiques bilingues à partir de corpus comparables sont basés sur une analyse du contexte lexical des mots et reposent sur la simple observation qu'un mot et sa traduction tendent à apparaître dans les mêmes contextes lexicaux. La mise en œuvre de cette observation repose sur l'identification d'*affinités du premier ordre* : « *Les affinités du premier ordre décrivent les mots qui sont susceptibles d'être trouvés dans le voisinage immédiat d'un mot donné.* [1] » (Grefenstette, 1994a, p. 279). Elles peuvent être représentées sous la forme d'un vecteur de contexte, où chaque élément du vecteur représente un mot qui apparaît dans différentes fenêtres contextuelles.

L'implémentation de l'approche standard peut être décrite par les quatre étapes suivantes (Rapp, 1995; Fung & McKeown, 1997) :

1. Identification des contextes lexicaux

Pour chaque partie du corpus comparable, le contexte de chaque mot plein i est extrait en repérant les mots qui apparaissent autour de lui dans une fenêtre contextuelle de n mots. Pour chaque mot i des langues source et cible, un vecteur de contexte $\mathbf{i}$ est ainsi obtenu. À chaque entrée $\mathbf{i}_j$ du vecteur est associée un score de cooccurrence des mots i et j. Habituellement, les mesures d'association comme l'information mutuelle (Fano, 1961), ou le taux de vraisemblance (Dunning, 1993) sont utilisées pour définir les entrées des vecteurs de contextes.

2. Transfert d'un mot à traduire

Les mots d'un vecteur de contexte i, pour lequel une traduction est recherchée, sont ensuite traduits en s'appuyant sur un dictionnaire bilingue. Si le dictionnaire propose plusieurs traductions pour un élément, nous ajoutons au vecteur de contexte de i l'ensemble des traductions proposées (lesquelles sont pondérées par la fréquence de la traduction en langue cible). Les entrées n'ayant pas de traductions dans le dictionnaire bilingue seront quant à elle tout simplement ignorées.

1. *First-order affinities describe what other words are likely to be found in the immediate vicinity of a given word*

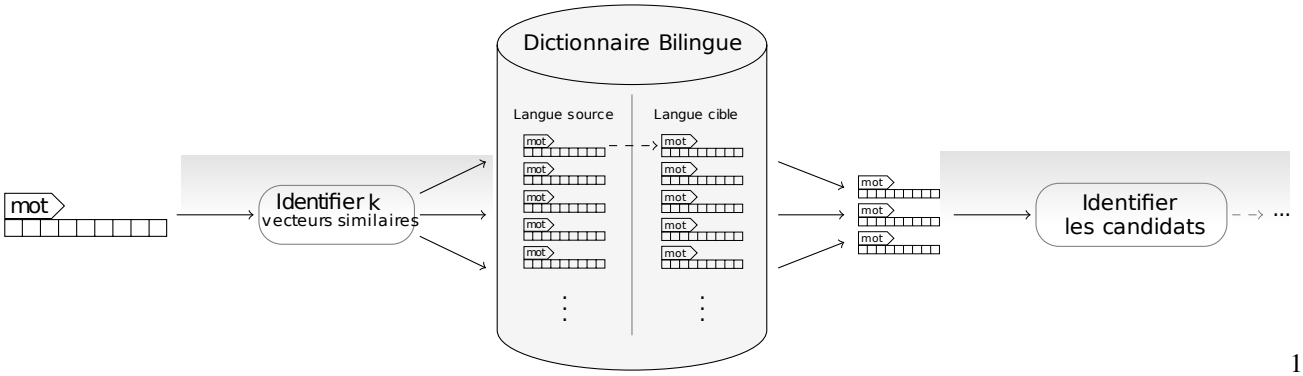
3. Identification des vecteurs proches du mot à traduire

Une mesure de similarité, $\mathrm{sim}(\bar{\mathbf{i}}, \mathbf{t})$, est utilisée pour calculer le score entre chaque mot, t, de la langue cible et le vecteur de contexte traduit du mot $\bar{\mathbf{i}}$. Parmi les mesures de similarité les plus souvent usitées pour cette tâche, nous retrouvons le cosinus (Salton & Lesk, 1968) ou le jaccard pondéré (Grefenstette, 1994b).

4. Obtention des traductions candidates

Les candidats à la traduction d'un mot i à traduire sont finalement ordonnés par ordre décroissant suivant leur score de similarité.

Deux remarques s'imposent ici en ce qui concerne cette approche standard. D'une part, cette approche met en œuvre différents paramètres (taille de la fenêtre contextuelle, choix des mesures d'association et de similarité...) dont il est parfois peu aisé d'identifier les valeurs adéquates pour une recherche optimum (voir par exemple le travail de (Laroche & Langlais, 2010) pour l'influence de ces paramètres sur la qualité des résultats). D'autre part, l'approche standard qui repose originellement sur des cooccurrences graphiques peut aussi être implémentée avec des cooccurrences syntaxiques (Yu & Tsujii, 2009; Otero, 2007).



1

FIGURE 1 – Illustration de la méthode par similarité interlangue.

2.2 Approche par similarité interlangue

Le principal inconvénient de l'approche standard est que ses performances dépendent grandement de la couverture du dictionnaire bilingue par rapport au corpus comparable. En effet, en traduisant un maximum d'entrées du vecteur de contexte du mot à traduire, on maximise les chances de retrouver sa traduction. Bien que la couverture du dictionnaire puisse être étendue en utilisant des dictionnaires spécialisés ou des thésaurus multilingues (Chiao & Zweigenbaum, 2003; Déjean *et al.*, 2002), la traduction des éléments du vecteur de contexte reste le cœur de cette approche.

Dans le but d'être moins dépendants de ce dictionnaire, (Déjean & Gaussier, 2002) ont proposé une extension de l'approche standard connue sous le nom d'approche par similarité interlangue. Cette approche se base sur l'idée que les mots ayant le même sens, partagent les mêmes environnements lexicaux. Elle repose sur l'identification d'*affinités du second ordre* : « *Les affinités du second ordre dévoilent quels mots partagent les mêmes environnements. Les mots partageant des affinités du second ordre n'ont pas besoin d'apparaître ensemble, mais leurs environnements sont semblables.* [2] »

Dans cette approche, le dictionnaire bilingue établit un pont entre les langues du corpus comparable. L'approche par similarité interlangue est basée sur ce principe et évite les traductions directes des éléments des vecteurs de contextes comme le montre la figure 1. L'implémentation de cette approche peut être réalisée en quatre étapes où la première et la dernière sont identiques à celles de l'approche standard (Déjean & Gaussier, 2002; Daille & Morin, 2005) :

2. Sélection des k plus proches voisins

Pour chaque mot à traduire, les k plus proches voisins sont identifiés parmi les entrées du dictionnaire selon

2. *Second-order affinities show which words share the same environments. Words sharing second-order affinities need never appear together themselves, but their environments are similar*

$\mathrm{sim}(\mathbf{i}, \mathbf{s})$. Chaque plus proche voisin est ensuite traduit à l'aide du dictionnaire bilingue, et le vecteur de contexte de langue cible $\overline{\mathbf{s}}$ correspondant à la traduction est sélectionné. Si pour un plus proche voisin il existe plusieurs traductions, $\overline{\mathbf{s}}$ est donné par l'union des vecteurs correspondant aux différentes traductions. Il est à noter que les vecteurs de contexte ne sont pas traduits directement, ce qui réduit l'influence du dictionnaire.

3. Identification des vecteurs proches du mot à traduire

La mesure de similarité, $\mathrm{sim}(\overline{\mathbf{s}}, \mathbf{t})$, est utilisée pour calculer le score entre chaque mot t de la langue cible en fonction des k plus proches voisins. Le score final attribué à chaque mot t est donné par :

$$\mathrm{sim}(\mathbf{i}, \mathbf{t}) = \sum_{s \in k\text{PPV}} \mathrm{sim}(\mathbf{i}, \mathbf{s}) \times \mathrm{sim}(\overline{\mathbf{s}}, \mathbf{t}) \quad (1)$$

Une autre manière de calculer le score de similarité a été proposée par (Daille & Morin, 2005). Les auteurs calculent alors le barycentre des vecteurs de contexte des k plus proches voisins.

3 Extraction lexicale bilingue par métarecherche

3.1 Motivations

L'approche proposée par (Déjean & Gaussier, 2002) introduit implicitement le problème du choix de la valeur adéquate de k dans les k plus proches voisins. D'une manière générale, la valeur optimale de k dépend des données mises en jeu. Cette valeur est souvent définie de façon empirique, bien qu'il soit possible de la déterminer par validation croisée. L'approche par similarité interlangue (ASI) appliquée à nos données s'est révélée très sensible vis-à-vis du paramètre k. En effet, pour des valeurs de k supérieures à 20, la précision chute de façon significative. De plus, il n'est pas possible de déterminer des intervalles de stabilité relative pour k. Le choix du paramètre k devient alors crucial.

En partant du principe que chaque mot contribue à la caractérisation du mot à traduire, notre proposition vise non seulement à améliorer la précision, mais aussi à être plus robuste vis-à-vis du nombre de plus proches voisins. En poussant l'analogie des approches inspirées de la RI (Fung & Lo, 1998) plus loin, nous proposons une nouvelle façon d'aborder le problème de l'extraction lexicale bilingue à partir de corpus comparables, en le considérant comme un problème de fusion de résultats analogue à celui rencontré par les métamoteurs de recherche.

L'objectif de la métarecherche est de fusionner les classements renvoyés par plusieurs systèmes de RI, en une liste unique, afin d'obtenir un système combiné qui soit plus performant que les systèmes individuels (Aslam & Montague, 2001). Puisque chacun des k plus proches voisins produit un classement différent, la métarecherche fourni un cadre adéquat pour exploiter l'information véhiculée par chacun des k classements. En outre, un intérêt particulier est donné aux mots candidats à la traduction d'un mot donné. En effet, partant du principe que les corpus contiennent beaucoup de bruit, il n'est pas rare de rencontrer des mots qui soient proches d'un nombre important de mots du dictionnaire, et ainsi, viennent parasiter le modèle et fausser les résultats. En effet, pour traduire un mot, le système choisit un nombre k de plus proches voisins en langue source, puis il cherche en langue cible les candidats les plus proches des traductions de ces k plus proches voisins sans tenir compte de la relation de ces candidats avec le reste des mots du dictionnaire. Pour pallier cela, nous construisons un modèle qui prend en compte cette information en accordant plus de confiance aux candidats qui sont plus proches des k plus proches voisins que du reste des entrées du dictionnaire.

3.2 Approche par métarecherche

Cette section décrit notre extension de l'approche par similarité interlangue. Les différents modes de fusion définis ici se basent sur les éléments décrits dans la table 1.

La première étape de notre méthode consiste à fixer le nombre de plus proches voisins d'un mot à traduire. La valeur de k est déterminée empiriquement. Cependant, intuitivement mais aussi à travers nos expériences, nous pouvons dire que la sélection d'un nombre faible de plus proches voisins est insuffisante dans la plus part des cas

Symbole	Définition
i	Le mot à traduire
t	Le mot candidat à la traduction de i
s	L'ensemble des plus proches voisins de i
$\overline{s}$	L'ensemble des traductions des plus proches voisins de i
k	Le nombre de plus proches voisins sélectionnés
n	L'ensemble de tous les voisins de t
u	Le nombre total de mots du dictionnaire
$occ_{\overline{s}(t)}$	L'effectif de t ie : avec combien de voisin t est-il en relation ?
$\mathrm{sim}(\overline{s_k}, t)$	Le score de similarité entre le k ième proche voisin de $\overline{s}$ et t
$\max_{\overline{s_k}}$	Le score maximum du candidat le plus proche de $\overline{s_k}$
$\max_{\overline{s}}$	Le score maximum du candidat le plus proche de l'ensemble $\overline{s}$
$\mathrm{sim}_k(s,t)$	Le score de similarité entre s et t par rapport au k ième plus proche voisin
$\mathrm{sim}(s,t)$	Le score de similarité entre s et t par rapport à l'ensemble des plus proches voisins
θ_t	Le paramètre de régulation ou facteur de confiance de t

TABLE 1 – Éléments de notation.

pour trouver la bonne traduction, et que la sélection d'un grand nombre de voisins, d'une part contredit la notion de plus proches voisins et d'autre part, induit la prise en compte de voisins éloignés qui peuvent fausser le modèle.

Une fois k fixé, nous considérons chaque liste de candidats renvoyée par un proche voisin indépendamment des autres. Ces candidats sont les mots dont les vecteurs de contexte sont les plus similaires au vecteur de contexte d'un voisin donnée. Dans nos expériences, la taille des listes a été fixée à 200. Partant du même principe que le choix du paramètre k. La taille de la liste joue un rôle important, en effet, une liste trop petite de candidats ne serait pas suffisante pour aider à trouver la bonne traduction, de la même façon, une liste trop importante de mots risque de rajouter du bruit car il faut garder en tête que les mots appartenant à une liste sont les traductions potentielles classées par ordre de score de similarité. Notre modèle privilégie les candidats qui apparaissent dans plusieurs listes, ainsi, plus l'effectif du mot candidat est important plus il a de chances d'être la bonne traduction, ceci dit, ceci reste valable si le candidat est bien classé, en revanche, s'il apparaît souvent mais en étant toujours mal classé par rapport aux différentes listes, ce mot a de fortes chances d'être une mauvaise traduction.

Dans l'approche par similarité interlangue, le calcul du score de similarité se fait sans prendre en considération les plus proches voisins d'une manière indépendante en amont, ainsi la fusion des scores est faite de telle sorte à ce que les candidats qui ont un score élevé par rapport à un proche voisin soit privilégiés par rapport à des candidats qui ont un score moins élevé mais qui apparaissent dans plusieurs listes. Notre approche vise à prendre en compte ce phénomène en normalisant les listes des plus proches voisins de la manière suivante :

$$\mathrm{sim}_k(i,t) = (\mathrm{sim}(i, s_k) \times \mathrm{sim}(\overline{s_k}, t)) \times \frac{\max_{\overline{s_k}}}{\max_{\overline{s}}} \tag{2}$$

Le raisonnement qui conduit à ce calcul est le suivant. Les scores des différents classements sont sur la même échelle car donnés par la même mesure de similarité. Ainsi, si $\max(l) \gg \max(m)$, cela veut dire que, selon le système, le classement l est plus « sûr » que le classement m (indépendamment de la réponse réelle).

Nous considérons ici que les classements, donnés par les k plus proches voisins du mot à traduire, sont le résultat de k moteurs de RI différents. À l'instar des métamoteurs de recherche, nous allons tenter de nous servir des scores des mots pour améliorer l'extraction bilingue.

Une des approches majeures en métarecherche est le modèle de fusion linéaire (LC) (Bartell *et al.*, 1994), où le score final d'un terme, i, est la somme pondérée de chacun des scores obtenus :

$$\mathrm{sim}(i,t) = \theta_t \times \frac{\sum_{j=1}^{k} \mathrm{sim}_j(i,t)}{\sum_{j=1}^{n} \mathrm{sim}(\overline{s_j}, t)} \tag{3}$$

Pour réduire l'influence des candidats à la traductions qui apparaissent dans différents contextes lexicaux et qui peuvent par leur forte fréquence d'apparition induire en erreur les systèmes d'extraction lexicale basés sur les

contextes, on se propose de prendre en compte ce phénomène en considérant en plus du score calculé à partir des k plus proches voisins, un score définit à partir de toutes les entrées du dictionnaire pour lequel le terme candidat est lié. L'équation 3 permet de calculer le score de similarité entre i et t en prenant en considération le score de similarité par rapport aux k plus proche voisins choisis, normalisé par la somme des scores de t par rapport à tout ses voisins pondéré par le paramètre de confiance θ. Le poids θ est donné par :

$$\theta_t = occ_{\overline{s}}(t) \times \frac{(u - (k - occ_{\overline{s}}(t)))}{(u - occ_n(t))} \tag{4}$$

L'équation 4 prend en compte l'effectif du candidat par rapport aux k plus proches voisins, c'est-à-dire, le nombre de voisins avec lesquels le mot t est en relation. Ceci est représenté par $occ_{\overline{s}}(t)$. Nous privilégions ainsi les mots avec un effectif élevé. Le numérateur $(u - (k - occ_{\overline{s}}(t)))$ permet de considérer l'effectif de t dans l'ensemble $\overline{s}$ par rapport à tous les mots du dictionnaire. On normalisera ensuite par la distribution de t par rapport à tous ses voisins à l'aide de $u - occ_n(t)$. Le paramètre θ permet donc d'accorder plus de confiance à un mot candidat à la traduction qui a un effectif élevé par rapports aux k plus proches voisins mais qui a aussi un effectif faible par rapport au reste de ses voisins.

4 Expériences et résultats

4.1 Ressources linguistiques

Dans le cadre de cette étude, nous avons construit un corpus comparable spécialisé français-anglais à partir de documents extraits du portail Elsevier [3]. L'ensemble des documents collectés relève du domaine médical restreint à la thématique du « cancer du sein ». Nous avons utilisé l'interface de recherche du portail pour sélectionner les publications scientifiques comportant dans le titre ou les mots clés le terme *cancer du sein* en français et *breast cancer* en anglais pour la période de 2001 à 2008. Les documents ont été nettoyés et normalisés à travers les traitements suivants : segmentation en occurrences de formes, étiquetage morpho-syntaxique et lemmatisation. Enfin, les mots agrammaticaux ont été supprimés et les mots apparaissant moins de deux fois dans la partie française et dans chaque partie anglaise écartés. Nous avons ainsi construit un corpus comparable spécialisé d'environ 1 million de mots qui est composé de 130 documents pour le français (7 376 mots distincts) et 103 documents pour l'anglais (8 457 mots distincts).

Le dictionnaire français-anglais nécessaire aux différentes approches comporte, après normalisation, 22 300 mots pour le français avec en moyenne 1,6 traductions par entrée. Il s'agit d'un dictionnaire de langue générale qui ne contient que peu de termes en rapport avec le domaine médical.

Pour évaluer les différentes approches utilisées dans cet article, nous avons sélectionné 400 couples de mots simples français-anglais à partir du meta-thesaurus UMLS [4] et du *Grand dictionnaire terminologique* [5]. Nous n'avons ensuite retenu que les couples pour lesquels le mot français apparaît au moins cinq fois dans la partie française et sa traduction au moins cinq fois dans la partie anglaise. Au terme de ce processus de sélection, nous disposons d'une liste de référence composée de 122 couples de termes simples français-anglais. Cette méthode de création d'une liste de référence est différente de celle proposée par (Déjean & Gaussier, 2002) qui construit sa liste à partir d'un sous ensemble du dictionnaire bilingue. Nous pensons que cette approche, plus fiable d'un point de vue statistique, ne correspond pas aux véritables difficultés rencontrées avec des corpus spécialisés. En effet, en domaine spécialisé les termes qui représentent une difficulté de traduction n'appartiennent par essence que rarement au dictionnaire de langue générale. En ce sens, nous préférons construire notre liste de référence en nous appuyant sur des nomenclatures attestées de termes du domaine non présent dans notre dictionnaire bilingue.

3. www.elsevier.com
4. www.nlm.nih.gov/research/umls
5. www.granddictionnaire.com

4.2 Paramètres expérimentaux

Trois paramètres communs à toutes les approches sont à fixer : i) la mesure d'association, ii) la mesure de similarité et iii) la taille de la fenêtre utilisée pour construire les vecteurs de contexte. Comme mesure de similarité nous avons choisit le jaccard pondéré (Grefenstette, 1994b) :

$$\mathrm{sim}(\mathbf{i},\mathbf{j}) = \frac{\sum_t \min(\mathbf{i}_t, \mathbf{j}_t)}{\sum_t \max(\mathbf{i}_t, \mathbf{j}_t)} \tag{5}$$

Les entrées du vecteur de contexte ont été déterminées par la mesure d'association du taux de vraisemblance (Dunning, 1993). La fenêtre contextuelle a été fixée à 7, partant de l'idée qu'elle approxime les dépendances syntaxiques. En plus de ces paramètres, notre approche ainsi que l'approche par similarité interlangue, ont besoin de définir le nombre de plus proche voisins.

Nous ne détaillons pas plus le choix de ces paramètres et renvoyons le lecteur vers (Morin, 2009) qui motive pour les mêmes ressources le choix de ces paramètres.

4.3 Résultats

Pour évaluer les performances de notre approche, nous utilisons comme référence l'approche par similarité interlangue (ASI) proposée par (Déjean & Gaussier, 2002). Nous comparons l'ASI avec les deux stratégies de l'approche par métarecherche définies dans la section 3 : i) le modèle qui se base sur les scores de similarité (AMS) sans tenir compte de la fiabilité des candidats ; ii) le modèle des sources multiples (AMF) qui prend en compte cette information. Nous allons étudier la stabilité des différentes stratégies de la méthode métarecherche en fonction de la variation des plus proches voisins.

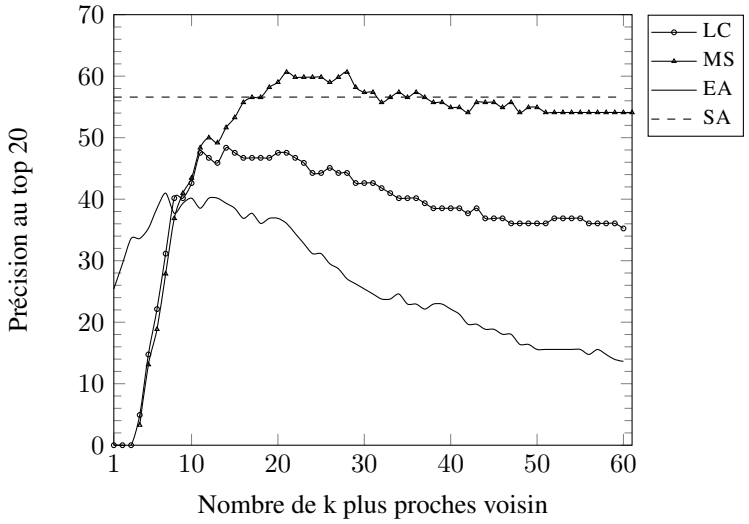


FIGURE 2 – Précision au top 20 en fonction du nombre de ppv.

La figure 2 montre la précision au top 20 en fonction de k. L'approche par similarité interlangue (ASI) atteint sa meilleure performance pour un $k = 7$ avec une précision de $40,98\%$, cette précision commence à décroitre d'une manière significative à partir de $k = 20$.

L'approche par métarecherche qui ne prend en considération que les scores de similarité (AMS) sans considérer la fiabilité des termes candidats à la traduction, montre de meilleurs résultats que la méthode de référence (ASI) à partir de $k = 10$ et obtient une précision maximale de $48,36\%$ pour un $k = 14$. On remarque aussi que la courbe correspondant au modèle AMS reste au-dessus de la méthode ASI malgré l'augmentation du paramètre k. La courbe correspondant au modèle de l'approche par métarecherche qui prend en compte la fiabilité des candidats

(AMF) est toujours au-dessus des autres à partir de $k = 10$. L'approche AMF améliore considérablement la précision et atteint sa meilleure performance avec $60,65\%$ pour un $k = 21$. Nous estimons que pour avoir une bonne exploitation des informations fournies par les différents k plus proches voisins en termes de score de similarité, notre système a besoin d'un minimum de k qui de par nos expériences semble être $k = 10$, ce qui explique les faibles résultats pour un $k < 10$. La raison des faibles résultats est simplement que notre système se base sur l'effectif des candidats à la traduction par rapport au paramètre k, en d'autres termes, de combien de proches voisins un candidat est-il proche ? il est évident qu'avec un k faible la notion d'effectif n'a pas assez de poids. Nous considérons aussi, que les candidats à la traduction proches d'un nombre très petit de voisins comme étant peu fiables. Ces candidats sont donc ignorés.

Nous pouvons noter à travers la figure 2 que les modèles AMF et AMS sont toujours meilleurs que la méthode de référence (ASI) (à partir de $k = 10$). De plus, ces modèles offrent une meilleure stabilité quant à la variation des k plus proches voisins. Quoique la précision décroisse en augmentant les valeurs de k, ceci se fait de manière moins rapide que l'approche de référence (ASI).

Nous comparons aussi nos résultats avec ceux obtenus par l'approche standard (AS). Celle ci est représentée dans la figure 2 par une droite car elle ne dépend pas du paramètre k. L'approche standard (AS) obtient une précision de $56,55\%$. Bien qu'elle soit au-dessus de l'approche par similarité interlangue (ASI) ainsi que du modèle AMS de l'approche par métarecherche, elle est en dessous de notre modèle AMF pour des valeurs de k entre 20 et 35. Nous pouvons ainsi considérer l'approche par métarecherche comme supérieure à l'approche de référence (ASI) mais aussi comme étant compétitive par rapport à l'approche standard (AS).

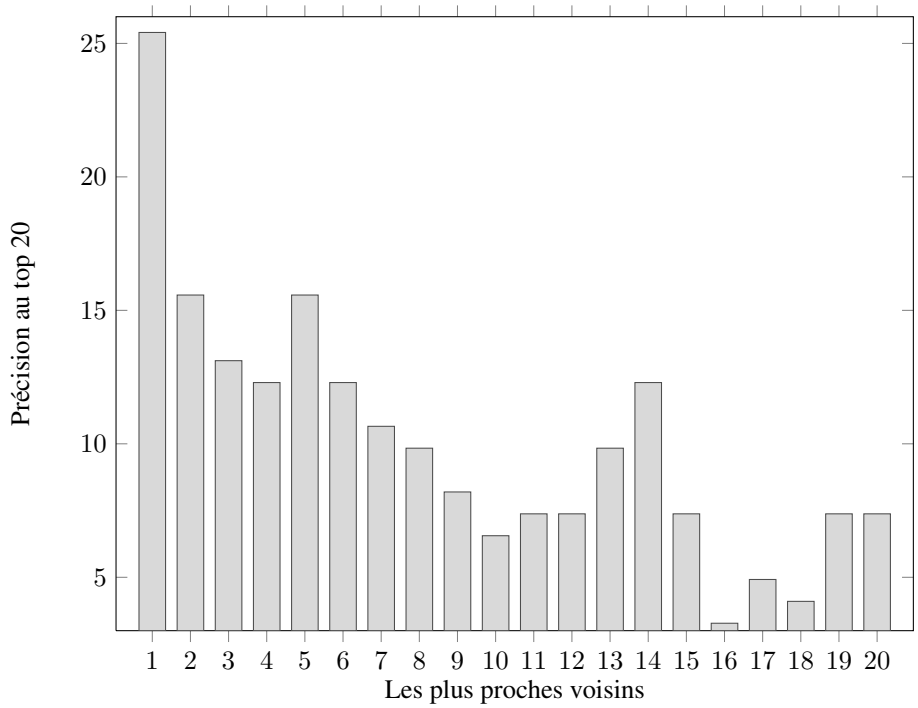


FIGURE 3 – Précision au top 20 pour chacun des 20-plus proches voisins . La précision est calculée en considérant les k plus proches voisins indépendamment les uns des autres.

La figure 3 montre la contribution de chaque plus proche voisin indépendamment des autres. Ceci confirme l'intuition que chaque proche voisin contribue à la caractérisation du mot à traduire, et confirme notre intuition sur le fait de les considérer indépendamment les uns des autres a priori et ceci en dressant pour chacun d'eux une liste de candidats, pour ensuite les combiner et ainsi améliorer les performances.

Il est à noter que les plus proches voisins sont ordonnés du plus proche voisin du mot à traduire au plus éloigné. Bien que chaque plus proche voisin ne puisse traduire qu'un nombre assez faible de mots, en utilisant l'idée de l'approche par métarecherche, nous pouvons améliorer les performances en termes de précision. Ainsi l'idée du

paradigme de notre méthode (AMF) est de prendre en compte l'information véhiculée par tous les plus proches voisins ainsi que le degré de fiabilité des candidats pour améliorer les performance du processus d'extraction lexicale.

Approches	Top 5	Top 10	Top 15	Top 20
AS	37,70%	45,08%	52,45%	56,55%
ASI	21,31%	31,14%	36,88%	40,98%
AMF	**40,98%**	**54,09%**	**56,55%**	**60,65%**

TABLE 2 – Précision aux tops 5, 10, 15, 20 des méthodes AS, ASI et AMF

Enfin comme dernier résultat, nous présentons dans le tableau 3 une comparaison des approches standard (AS), par similarité interlangue (ASI) et par méta-recherche (AMF), pour le top 5, 10, 15 et 20 et ceci en choisissant la meilleure configuration des paramètres de chaque approche. Nous constatons que notre approche AMF obtient une meilleure précision dans chaque situation. Partant du principe que les systèmes d'extraction lexicale tentent d'approcher le top 10 voir le top 5, nous considèreront nos résultats comme étant encourageants notamment pour le top 10 où AMF atteint $54,09\%$ ce qui n'est pas loin de l'approche standard au top 20.

4.4 Discussion

Les approches par similarité interlangue (ASI) et par métarecherche se basent sur les k plus proches voisins pour identifier les meilleurs candidats à la traduction. L'approche ASI effectue une fusion en amont, privilégiant ainsi une vue globale des k plus proches voisins. Ceci peut se révéler problématique, car une bonne traduction pourrait être noyée dans la masse, et ainsi être écartée de la liste des candidats, si des mots plus fréquents viennent à apparaître dans le contexte du mot à traduire. Plus précisément, si des mots obtiennent des scores de similarité très élevés par rapport à un seul plus proche voisin et que d'autres obtiennent des scores moins élevés mais proches de plusieurs plus proches voisins du mot à traduire, ces mots seront moins bien classés voir mal classés. Pour pallier ce problème, l'approche AMF considère dans un premier temps, chaque plus proche voisin comme étant une source d'information indépendante des autres, privilégiant ainsi sa liste de candidats en fixant une taille arbitraire (généralement autour de 200 dans nos expériences), pour ensuite effectuer une fusion en aval des plus proches voisins après avoir normalisé les scores de similarité comme décrit en section 3. En outre, l'approche AMF introduit une mesure de fiabilité, en considérant les plus proches voisins des mots candidats à la traduction comme étant proches des k plus proches voisins du mot à traduire, ainsi que tous les voisins de ces candidats, pour éloigner des mots qui apparaitraient trop fréquemment et dans trop de contextes. Car ne l'oublions pas, ces approches se basent uniquement sur une représentation graphique des données qui induit un certain volume de bruit, lequel serait sans doute mieux traité par une analyse plus fine du contexte. Il est évident que plus les mots sont fréquents dans le corpus plus on a une représentation riche de leur contexte. Cette remarque nous amène à nous interroger sur les fréquences des k plus proches voisins du mot candidat à la traduction. En effet, si un plus proche voisin apparaît fréquemment en langue source et que sa traduction en langue cible est faible ou inversement, quel serait l'impact de ce déséquilibre sur les résultats ? Aucune étude à notre connaissance n'a approfondi ce sujet. Quoique rien ne nous permette d'affirmer une quelconque relation entre ce déséquilibre et une éventuelle traduction erronée, nous pouvons néanmoins supposer que cela est nuisible à une représentation riche du contexte du mot, car un mot peu fréquent apporte moins d'information qu'un mot très fréquent et ceci toujours en se basant sur l'idée de la coloration graphique qui caractérise le contexte. Ainsi nous nous attellerons dans nos travaux futurs à étudier cette problématique. Enfin, les plus proches voisins ont été fixés d'une manière empirique dans les deux approches ASI et AMF, et dans toutes les évaluations. Nous avons fixé un même k pour tous les mots de la liste d'évaluation. L'état de l'art ne spécifie aucune manière efficace de choisir ce paramètre k. Néanmoins, nous sommes en droit de nous interroger pour savoir s'il existe un nombre k idéal de plus proches voisins qui puisse garantir une bonne traduction de tous les mots de la liste d'évaluation ? On serait plutôt tenté de dire qu'il existe un k pour chaque mot à traduire mais que celui-ci varie selon les mots. Là encore, nos travaux futurs devront répondre à cette question clé.

5 Conclusion

Nous avons présenté dans cet article une nouvelle manière d'aborder le problème de l'extraction lexicale bilingue à partir de corpus comparables en nous appuyant sur le principe des métamoteurs de recherche. Nous avons ainsi présenté une nouvelle approche simple et robuste qui revisite la méthode par similarité interlangue pour présenter un modèle inspiré par les métamoteurs de recherche d'information. Ce modèle qui prend en compte la distribution des candidats à la traduction non seulement par rapport au k plus proches voisins du mot à traduire mais aussi par rapport à tout leurs voisins, a permis un gain significatif en terme de précision. Les résultats empiriques que nous obtenons montrent que les performances de ce nouveau modèle sont toujours supérieures à celles obtenues avec l'approche par similarité interlangue pour $k > 10$, mais aussi comme étant compétitives par rapport à l'approche standard.

Remerciements

Ce travail qui s'inscrit dans le cadre du projet METRICC (`www.metricc.com`) a bénéficié d'une aide de l'Agence Nationale de la Recherche portant la référence ANR-08-CORD-009.

Références

ASLAM J. A. & MONTAGUE M. (2001). Models for Metasearch. In *SIGIR '01, proceedings of the 24th Annual SIGIR Conference*, p. 276–284.

BARTELL B. T., COTTRELL G. W. & BELEW R. K. (1994). Automatic combination of multiple ranked retrieval systems. In *SIGIR '94, proceedings of the 17th Annual SIGIR Conference*, p. 173–181.

CHIAO Y.-C. & ZWEIGENBAUM P. (2003). The Effect of a General Lexicon in Corpus-Based Identification of French-English Medical Word Translations. In R. BAUD, M. FIESCHI, P. LE BEUX & P. RUCH, Eds., *The New Navigators : from Professionals to Patients, Actes Medical Informatics Europe*, volume 95 of *Studies in Health Technology and Informatics*, p. 397–402, Amsterdam : IOS Press.

DAILLE B. & MORIN E. (2005). French-English Terminology Extraction from Comparable Corpora. In *Proceedings of the 2nd International Joint Conference on Natural Language Processing (IJCLNP'05)*, p. 707–718, Jeju Island, Korea.

DÉJEAN H. & GAUSSIER E. (2002). Une nouvelle approche à l'extraction de lexiques bilingues à partir de corpus comparables. *Lexicometrica, Alignement lexical dans les corpus multilingues*, p. 1–22.

DÉJEAN H., SADAT F. & GAUSSIER E. (2002). An approach based on multilingual thesauri and model combination for bilingual lexicon extraction. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING'02)*, p. 218–224, Tapei, Taiwan.

DUNNING T. (1993). Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics*, **19**(1), 61–74.

FANO R. M. (1961). *Transmission of Information : A Statistical Theory of Communications*. Cambridge, MA, USA : MIT Press.

FUNG P. (1998). A Statistical View on Bilingual Lexicon Extraction : From ParallelCorpora to Non-parallel Corpora. In D. FARWELL, L. GERBER & E. HOVY, Eds., *Proceedings of the 3rd Conference of the Association for Machine Translation in the Americas (AMTA'98)*, p. 1–16, Langhorne, PA, USA.

FUNG P. & LO Y. Y. (1998). An ir approach for translating new words from nonparallel, comparable texts. In *Proceedings of the 17th international conference on Computational linguistics (COLING'98)*, p. 414–420.

FUNG P. & MCKEOWN K. (1997). Finding Terminology Translations from Non-parallel Corpora. In *Proceedings of the 5th Annual Workshop on Very Large Corpora (VLC'97)*, p. 192–202, Hong Kong.

GREFENSTETTE G. (1994a). Corpus-Derived First, Second and Third-Order Word Affinities. In *Proceedings of the 6th Congress of the European Association for Lexicography (EURALEX'94)*, p. 279–290, Amsterdam, The Netherlands.

GREFENSTETTE G. (1994b). *Explorations in Automatic Thesaurus Discovery*. Boston, MA, USA : Kluwer Academic Publisher.

LAROCHE A. & LANGLAIS P. (2010). Revisiting context-based projection methods for term-translation spotting in comparable corpora. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, p. 617–625, Stroudsburg, Pekin, Chine : Association for Computational Linguistics.

MORIN E. (2009). Apport d'un corpus comparable déséquilibré à l'extraction de lexiques bilingues. In *Actes de la 16ème Conférence annuelle sur le Traitement Automatique des Langues Naturelles (TALN) Senlis France.*, p. 101–110.

OTERO P. G. (2007). Learning bilingual lexicons from comparable english and spanish corpora. In *Proceedings of Machine Translation Summit XI*, p. 191–198.

RAPP R. (1995). Identify Word Translations in Non-Parallel Texts. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (ACL'95)*, p. 320–322, Boston, MA, USA.

SALTON G. & LESK M. E. (1968). Computer evaluation of indexing and text processing. *Journal of the Association for Computational Machinery*, **15**(1), 8–36.

YU K. & TSUJII J. (2009). Bilingual dictionary extraction from wikipedia. In *Proceedings of Machine Translation Summit XII*.